STATISTICS IN MEDICINE
Statist. Med. (in press)
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2321

Assessment on homogeneity tests for kappa statistics under equal prevalence across studies in reliability‡

Jun-mo Nam*,†

Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Department of Health & Human Services, Executive Plaza South, Room 8028, 6120 Executive Boulevard, MSC 7240, Rockville, MD 20892-7240, U.S.A.

SUMMARY

In this paper, we assess the performance of homogeneity tests for two or more kappa statistics when prevalence rates across reliability studies are assumed to be equal. The likelihood score method and the chi-square goodness-of-fit (GOF) test provide type 1 error rates that are satisfactorily close to the nominal level, but a Fleiss-like test is not satisfactory for small or moderate sample sizes. Simulations show that the score test is more powerful than the chi-square GOF test and the approximate sample size required for a specific power of the former is substantially smaller than the latter. In addition, the score test is robust to deviations from the equal prevalence assumption, while the GOF test is highly sensitive and it may give a grossly misleading type 1 error rate when the assumption of equal prevalence is violated. We conclude that the homogeneity score test is the preferred method. Published in 2005 by John Wiley & Sons, Ltd.

KEY WORDS: distortion of *p*-value; equal prevalence; homogeneity of kappas; reliability; sample size

1. INTRODUCTION

The intraclass kappa coefficient has been commonly applied to assess the reliability of the binary classification of a subject. For example, two raters independently classify a subject according to the presence or absence of a certain characteristic or a single rater blindly rates a subject twice as the positive or negative. In both cases, the intraclass kappa statistic can be used as a measure of agreement of two ratings when there is an assessment of the reliability of the ratings. Cohen's kappa [1] assumes that the probability of a positive by the first rating and that of the second one are different while the intraclass version of kappa assumes that

*Correspondence to: Jun-mo Nam, Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Department of Health & Human Services, Executive Plaza South, Room 8028, 6120 Executive Boulevard, MSC 7240, Rockville, MD 20892-7240, U.S.A.
†E-mail: namj@mail.nih.gov

Received 30 May 2003
Published in 2005 by John Wiley & Sons, Ltd. *Accepted 26 May 2005*

the two probabilities are equal. Under the model related to the intraclass kappa, successive $m(\geqslant 2)$ ratings per subject are interchangeable. In this report, we limit our attention to the intraclass kappa which is identical to the Scott index [2] and consider $m=2$ which is the most common case. For a measure of reliability, it should be noted that the intraclass form of kappa is applied but Cohen's kappa is not generally used, e.g. Reference [3]. Cohen's kappa is a special case of a weighted kappa when the costs of false positives and false negatives are same, e.g. Reference [4]. The intraclass version of kappa has also been discussed by, e.g. Dunn [5]. The intraclass kappa coefficient is algebraically equal to the inbreeding coefficient in population genetics, e.g. Reference [6].

Statistical methods involving the intraclass version of kappa agreement for a single reliability study have been well documented by many authors, e.g. References [3, 7–9] for estimation, and References [7, 10] for significance testing. Recently, there has been increased interest in comparing two or more kappa statistics in multiple studies or, in a stratified study, e.g. References [11–13]. Almost all statistical tests for the homogeneity of kappas were considered without assuming equal prevalences across studies. For some cases, it may be reasonable to assume equal prevalence based on previous studies or theoretical consideration. However, homogeneity tests for several kappas under the assumption of equal prevalences have not been fully studied. It could be interesting to examine the values of these tests against those values under unequal prevalence, e.g. Reference [12], when prevalences are equal.

In this paper, we investigate the statistical properties of homogeneity tests derived assuming equal prevalence across studies. In Section 3, Fleiss-like procedure, chi-square goodness-of-fit (GOF) and likelihood score tests for homogeneity of kappas based on equal prevalence are given along with their sample size requirements. In Section 4, these homogeneity tests are compared in terms of empirical type 1 error rate and power. In Section 5, the distortions of the level of significance of the tests are examined when the assumption of equal prevalence is violated. Sections 6 and 7 contain an example and concluding remarks.

2. NOTATION

Consider J independent studies involving n_j subjects for $j=1,2,\ldots,J$. For each study, a subject is blindly rated twice by an examiner with a rating of either positive or negative. Let the probabilities of a positive and a negative rating on a subject in the jth study be $\Pr(+)=p_j$ and $\Pr(-)=q_j$ where $p_j+q_j=1$ for $j=1,2,\ldots,J$. The n_j pairs of ratings can be distributed into three categories: $(+,+)$; $(+,-)$ or $(-,+)$; and $(-,-)$. The observed numbers of pairs in the three categories are x_{2j}, x_{1j} and x_{0j} and their corresponding probabilities are P_{2j}, P_{1j} and P_{0j} where the first subscript represents the number of positive ratings in a pair. Note that $x_{2j}+x_{1j}+x_{0j}=n_j$ and $P_{2j}+P_{1j}+P_{0j}=1$. The kappa coefficient, denoted κ_j, is the correlation coefficient between two ratings in a pair, and $\kappa_j=(P_{2j}-p_j^2)/(p_jq_j)=(P_{0j}-q_j^2)/(p_jq_j)$ which yields the multinomial probabilities: $P_{2j}(\kappa_j, p_j)=p_j^2+p_jq_j\kappa_j, P_{1j}(\kappa_j, p_j)=2p_jq_j(1-\kappa_j)$ and $P_{0j}(\kappa_j, p_j)=q_j^2+p_jq_j\kappa_j$ [3, 14] for $j=1,2,\ldots,J$. The notations for the observed data are summarized in Table I. The intraclass kappa, κ_j, is identical to the kappa by the standard definition, i.e. $\kappa_j=(p_{oj}-p_{ej})/(1-p_{ej})$ where $p_{oj}=P_{2j}+P_{0j}$ and $p_{ej}=p_j^2+q_j^2$.

Table I. Distribution of ratings in J different studies.

	Observed frequency				
Category	Study 1	Study 2	...	Study J	Sum
$(+,+)$	x_{21}	x_{22}	...	x_{2J}	$x_{2.}$
$(+,-)$ or $(-,+)$	x_{11}	x_{12}	...	x_{1J}	$x_{1.}$
$(-,-)$	x_{01}	x_{02}	...	x_{0J}	$x_{0.}$
Sum	n_1	n_2	...	n_J	$n_.$

3. HOMOGENEITY TEST AND REQUIRED SAMPLE SIZE

From the joint distribution of $x' = (x'_1, x'_2, \ldots, x'_J)$ where $x'_j = (x_{2j}, x_{1j}, x_{0j})$ for $j = 1, 2, \ldots, J$, the log likelihood is expressed as $\ln L(\boldsymbol{\kappa}, \mathbf{p}) = \sum_{j=1}^{J} \ln L_j(\kappa_j, p_j)$, where $\ln L_j(\kappa_j, p_j) = x_{2j} \cdot \ln\{p_j(p_j + q_j\kappa_j)\} + x_{1j} \cdot \ln\{2p_jq_j(1-\kappa_j)\} + x_{0j} \cdot \ln\{q_j(q_j + p_j\kappa_j)\}$ for $j = 1, 2, \ldots, J$. For the jth study, the MLEs of κ_j and p_j are $\hat{\kappa}_j = (4x_{2j}x_{0j} - x_{1j}^2)/\{(2x_{2j} + x_{1j})(2x_{0j} + x_{1j})\}$ and $\hat{p}_j = (2x_{2j} + x_{1j})/(2n_j)$ and the variance of $\hat{\kappa}_j$ is $\text{var}(\hat{\kappa}_j) = (1-\kappa_j)\{(1-\kappa_j)(1-2\kappa_j) + \kappa_j(2-\kappa_j)/(2p_jq_j)\}/n_j$, e.g. Reference [3]. Consider the comparison of kappa statistics from J independent reliability studies when the prevalences (i.e. probabilities of positive rating) are the same across the studies. Under the null hypothesis $H_0 : \kappa_j = \kappa$ for $j = 1, 2, \ldots, J$ assuming equal prevalence, i.e. $p_j = p$ for every j, the MLEs of κ and p are $\hat{\kappa} = (4x_{2.}x_{0.} - x_{1.}^2)/\{(2x_{2.} + x_{1.})(2x_{0.} + x_{1.})\}$ and $\hat{p} = (2x_{2.} + x_{1.})/(2n_.)$ where $x_{i.} = \sum_{j=1}^{J} x_{ij}$ for $i = 0, 1$ and 2, and $n_. = \sum_{j=1}^{J} n_{j.}$.

3.1. Fleiss-like method

The Fleiss-like [11] method for testing homogeneity of several kappa statistics assuming equal prevalence is

$$X_F^2 = \sum_{j=1}^{J} \omega_j(\hat{\kappa}, \hat{p}).(\hat{\kappa}_j - \hat{\kappa})^2 = \sum_{j=1}^{J} n_j\,(\hat{\kappa}_j - \hat{\kappa})^2 \big/ v(\hat{\kappa}, \hat{p}) \tag{1}$$

where $\omega_j(\hat{\kappa}, \hat{p})$ is the inverse of var $(\hat{\kappa}_j)$ evaluated at $\kappa_j = \hat{\kappa}$ and $p_j = \hat{p}$, i.e. $\omega_j(\hat{\kappa}, \hat{p}) = n_j/v$ where $v = (1-\hat{\kappa})^2[(1-2\hat{\kappa}) + \hat{\kappa}(2-\hat{\kappa})/\{2\hat{p}\hat{q}(1-\hat{\kappa})\}]$ for $j = 1, 2, \ldots, J$. The statistic, (1), is asymptotically distributed as a chi-square with $J-1$ degrees of freedom under $H_0 : \kappa_j = \kappa$ for $j = 1, 2, \ldots, J$. We reject H_0 at α when $X_F^2 \geqslant \chi^2_{J-1,\alpha}$, where $\chi^2_{J-1,\alpha}$ is the $100 \times (1-\alpha)$ percentile point of the chi-square distribution with $J-1$ degrees of freedom. Let $\bar{\kappa} = \sum_{j=1}^{J} t_j\kappa_j$, where t_j's are design fractions, i.e. $t_j = n_j/n.$ for $j = 1, 2, \ldots, J$ and $c = 1 - 2\bar{\kappa} + \bar{\kappa}(2-\bar{\kappa})/\{2pq(1-\bar{\kappa})\}$. Under the alternative $H_1 : \kappa_j \neq \kappa$ for any j, statistic (1) is asymptotically distributed as a non-central chi-square with $J-1$ degrees of freedom and non-centrality parameter which is $\lambda = \sum n_j(\kappa_j - \bar{\kappa})^2/\{c \cdot (1-\bar{\kappa})^2\}$ using, e.g. Reference [15]. The total sample size required for power $= 1-\beta$ of the Fleiss-like test at level α is obtained by $n_. = \lambda(J-1, 1-\beta, \alpha) \cdot c \cdot (1-\bar{\kappa})^2/\{\sum_{j=1}^{J} t_j(\kappa_j - \bar{\kappa})^2\}$ where $\lambda(J-1, 1-\beta, \alpha)$ is the value of the non-centrality parameter of the cumulative non-central chi-square distribution corresponding to power $= 1-\beta$ at level α, e.g. $\lambda\ (1, 0.8, 0.05) = 7.804$ for $J = 2$, 80 per cent power and 5 per cent level from tables

of the cumulative non-central chi-square distribution [16]. Note that the Fleiss-like method is based on assuming a normal distribution for the intraclass kappas, but the distribution of a kappa statistic converges quite slowly to the normal and the method may not be satisfactory for small or medium sample sizes.

3.2. Pearson's chi-square GOF test

Applying the Pearson's chi-square GOF test, we may test the homogeneity of kappa statistics assuming $p_j = p$ for every j using $X_G^2 = \sum_{i=0}^{2} \sum_{j=1}^{J} \{x_{ij} - n_j \cdot P_{ij}(\hat{\kappa}, \hat{p})\}^2 / \{n_j \cdot P_{ij}(\hat{\kappa}, \hat{p})\}$ where $P_{2j}(\hat{\kappa}, \hat{p}) = \hat{p}(\hat{p} + \hat{q}\hat{\kappa}), P_{1j}(\hat{\kappa}, \hat{p}) = 2\hat{p}\hat{q}(1 - \hat{\kappa})$ and $P_{0j}(\hat{\kappa}, \hat{p}) = \hat{q}(\hat{q} + \hat{p}\hat{\kappa})$ with $\hat{q} = 1 - \hat{p}$. The GOF statistic can be rewritten as

$$X_G^2 = \frac{\sum_{j=1}^{J} x_{2j}^2 / n_j}{\hat{p}(\hat{p} + \hat{q}\hat{\kappa})} + \frac{\sum_{j=1}^{J} x_{1j}^2 / n_j}{2\hat{p}\hat{q}(1 - \hat{\kappa})} + \frac{\sum_{j=1}^{J} x_{0j}^2 / n_j}{\hat{q}(\hat{q} + \hat{p}\hat{\kappa})} - n. \tag{2}$$

Henceforth, the GOF statistic, X_G^2, refers to the Pearson's chi-square GOF statistic under equal prevalence in this paper. Statistic (2) is approximately a chi-square with $2\ (J - 1)$ degrees of freedom under H_0. Under the alternative H_1, the GOF statistics is asymptotically a non-central chi-square with $2\ (J - 1)$ degrees of freedom and the non-centrality parameter as

$$\lambda_G = \frac{pq}{(1 - \bar{\kappa})} \left\{ 1 + \frac{1}{\bar{\kappa} + (1 - \bar{\kappa})^2} \right\} \cdot \left\{ \sum_{j=1}^{J} n_j (\kappa_j - \bar{\kappa})^2 \right\}$$

From λ_G, we have the total sample size for a required power of the test at α as

$$n_\cdot = \lambda(2J - 2, 1 - \beta, \alpha) \cdot (1 - \bar{\kappa}) \Bigg/ \left[pq \left\{ 1 + \frac{1}{\bar{\kappa} + (1 - \bar{\kappa})^2} \right\} \cdot \sum_{j=1}^{J} t_j (\kappa_j - \bar{\kappa})^2 \right]$$

Note that Donner *et al.* [12] presented a GOF test of homogeneity of kappas without the assumption of equal prevalence. We may call it as the Donner's GOF test.

3.3. Likelihood score method

Denote partial derivatives as $S_j(\kappa_j, p) \equiv \partial \ln L_j(\kappa_j, p) / \partial \kappa_j$ for $j = 1, 2, \ldots, J$. The score and its variance evaluated at $\kappa_j = \hat{\kappa}$ and $p = \hat{p}$ are $S_j(\hat{\kappa}, \hat{p}) = \{x_{2j}/(\hat{p} + \hat{q}\hat{\kappa}) + x_{0j}/(\hat{q} + \hat{p}\hat{\kappa}) - n_j\}/(1 - \hat{\kappa})$ (see Appendix A) and $\omega_j(\hat{\kappa}, \hat{p}) = n_j / v$ for $j = 1, 2, \ldots, J$. The score statistic,

$$X_s^2 = \sum_{j=1}^{J} \{S_j(\hat{\kappa}, \hat{p})\}^2 / \omega_j(\hat{\kappa}, \hat{p}) \tag{3}$$

is asymptotically distributed as a chi-square with $J - 1$ degrees of freedom using a general theory by, e.g. References [17, 18]. We reject the homogeneity of the J kappa statistics when (3) is larger than a critical value. Under H_1, the homogeneity score test is a non-central

chi-square distribution with $J-1$ degrees of freedom and non-centrality parameter

$$\lambda_s = \bar{c}\left\{\frac{(1+\bar{\kappa})pq}{\bar{\kappa}+(1-\bar{\kappa})pq}\right\}^2 \cdot \left\{\sum_{j=1}^{J} n_j(\kappa_j-\bar{\kappa})^2\right\}$$

The sample size required for power $=1-\beta$ of the score test at level α [13] is

$$n_\cdot = \lambda(J-1, 1-\beta, \alpha) \Bigg/ \left[\bar{c}\left\{\frac{(1+\bar{\kappa})pq}{\bar{\kappa}+(1-\bar{\kappa})^2 pq}\right\}^2 \cdot \left\{\sum_{j=1}^{J} t_j(\kappa_j-\bar{\kappa})^2\right\}\right]$$

4. NUMERICAL EVALUATION

To investigate the actual level and power of the three homogeneity tests for small or moderate sample sizes, we conducted Monte Carlo experiments with 1000 simulations for various values of p (prevalence) and a wide range of kappa coefficients considering two reliability studies ($J=2$). For calculations of type 1 error rates of the homogeneity tests in simulations, we excluded those sampling points where a statistic is undefined. Empirical error rates are those conditional on both intraclass kappas being well-defined. Note that the probability of undefined cases is very small or near to zero in this study. Results of simulations are summarized in Tables II and III. Table II shows that the empirical type 1 error rates of the GOF test and score method for testing homogeneity of kappa statistics under the assumption of equal prevalence are reasonably close to the nominal level. The Fleiss-like test tended to be anticonservative for

Table II. Empirical type 1 error rates of homogeneity tests at $\alpha=0.05$ for $\kappa_1=\kappa_2=\kappa$ under the correct assumption of $p_1=p_2=p$ based on 1000 simulations.

		$n_1=20, n_2=30$				$n_1=40, n_2=60$			
p	κ	X_F^2	X_D^2	X_G^2	X_s^2	X_F^2	X_D^2	X_G^2	X_s^2
0.25	0.2	0.074	0.062	0.048	0.056	0.054	0.051	0.038	0.046
	0.4	0.075	0.065	0.052	0.055	0.054	0.058	0.042	0.059
	0.6	0.068	0.057	0.048	0.053	0.051	0.047	0.045	0.053
	0.8	0.074	0.075	0.046	0.049	0.054	0.056	0.045	0.050
0.35	0.2	0.047	0.049	0.047	0.042	0.042	0.042	0.043	0.040
	0.4	0.070	0.063	0.057	0.058	0.060	0.059	0.061	0.050
	0.6	0.073	0.058	0.055	0.062	0.060	0.057	0.052	0.056
	0.8	0.063	0.059	0.045	0.049	0.071	0.070	0.053	0.066
0.40	0.2	0.060	0.057	0.053	0.050	0.064	0.065	0.054	0.058
	0.4	0.057	0.052	0.055	0.050	0.052	0.052	0.050	0.046
	0.6	0.065	0.058	0.058	0.058	0.069	0.065	0.059	0.066
	0.8	0.044	0.039	0.048	0.035	0.063	0.057	0.062	0.056
0.50	0.2	0.054	0.051	0.055	0.048	0.061	0.058	0.068	0.056
	0.4	0.051	0.043	0.055	0.040	0.062	0.061	0.057	0.058
	0.6	0.051	0.044	0.036	0.041	0.054	0.053	0.049	0.053
	0.8	0.048	0.046	0.054	0.039	0.059	0.056	0.053	0.058

X_F^2, X_G^2 and X_s^2 refer to the Fleiss-like (1), chi-square GOF (2) and score (3) tests. X_D refers to the chi-square GOF method without assuming equal prevalence by Donner *et al.* [12].

Table III. Empirical powers of homogeneity tests at $\alpha = 0.05$ under the correct assumption of $p_1 = p_2$ based on 1000 simulations (when $K_1 = K_2$, level is estimated and when $k_1 \neq k_2$, power is estimated).

			$n_1 = 20, n_2 = 30$				$n_1 = 40, n_2 = 60$				$n_1 = 100, n_2 = 150$			
p	k_1	k_2	X_F^2	X_D^2	X_G^2	X_s^2	X_F^2	X_D^2	X_G^2	X_s^2	X_F^2	X_D^2	X_G^2	X_s^2
0.2	0.2	0.2	0.074	0.063	0.058	0.053	0.075	0.074	0.055	0.070	0.056	0.055	0.053	0.054
		0.4	0.137	0.124	0.082	0.113	0.178	0.177	0.121	0.165	0.237	0.268	0.199	0.267
		0.6	0.297	0.280	0.196	0.254	0.464	0.466	0.357	0.451	0.788	0.798	0.708	0.798
		0.8	0.590	0.570	0.425	0.547	0.835	0.846	0.728	0.842	0.994	0.994	0.988	0.995
	0.4	0.4	0.088	0.065	0.051	0.050	0.070	0.069	0.048	0.063	0.054	0.054	0.060	0.053
		0.6	0.166	0.145	0.092	0.125	0.189	0.189	0.135	0.180	0.314	0.318	0.245	0.328
		0.8	0.363	0.339	0.240	0.324	0.544	0.543	0.413	0.535	0.893	0.895	0.837	0.903
0.4	0.2	0.2	0.060	0.057	0.058	0.050	0.064	0.065	0.054	0.058	0.054	0.054	0.056	0.052
		0.4	0.143	0.139	0.097	0.129	0.190	0.188	0.138	0.185	0.367	0.367	0.279	0.364
		0.6	0.354	0.349	0.267	0.330	0.601	0.601	0.498	0.594	0.916	0.915	0.848	0.918
		0.8	0.711	0.701	0.595	0.683	0.945	0.943	0.900	0.938	1.000	1.000	1.000	1.000
	0.4	0.4	0.057	0.052	0.056	0.050	0.052	0.052	0.050	0.046	0.062	0.061	0.066	0.060
		0.6	0.157	0.153	0.128	0.141	0.242	0.240	0.178	0.226	0.428	0.428	0.336	0.430
		0.8	0.451	0.440	0.325	0.424	0.706	0.700	0.602	0.698	0.971	0.971	0.946	0.974

X_F^2, X_G^2 and X_s refer to the Fleiss-like (1), chi-square GOF (2) and score (3) tests. X_D^2 refers to the chi-square GOF method without assuming equal prevalence by Donner *et al.* [12].

a small sample size. Note that results for $p = 0.65$ and 0.75 are similar to those for $p = 0.35$ and 0.25, respectively. This simulation study showed that the GOF and score tests for testing homogeneity of kappa statistics based on the asymptotic theory can be used for small or moderate sample sizes. Table III indicates the Fleiss-like and score tests were consistently more powerful than the GOF test. The empirical power of the Fleiss-like test was inflated when compared with the GOF and score tests since the type 1 error rate of the Fleiss-like test was greater than the others. Donner *et al.* [12] proposed the GOF procedure and Fleiss-like method without assuming equal prevalence. Table II indicated that the actual level of the Donner's GOF procedure was generally greater than the nominal level for small and moderate sample sizes when prevalences are the same. The range of empirical type 1 error rates of the Fleiss-like method without assuming equal prevalence is large, particularly, for a small sample size and the Fleiss-like test may provide a highly unreliable level of significance. For large sample sizes in Table III, the Fleiss-like, Donner's GOF and score tests were similar and the GOF test was inferior to the above three tests in power under the equal prevalence.

5. DISTORTION OF LEVEL OF SIGNIFICANCE OF TEST

To examine the behaviour of the level of significance of the homogeneity tests when prevalences are unequal, we carried out simulations for $(p_1, p_2) = (0.2, 0.3), (0.2, 0.5), (0.3, 0.5), (0.2, 0.8)$ and $(0.8, 0.2)$, $\kappa = 0.2, 0.4, 0.6, 0.8$ and $(n_1, n_2) = (20, 30), (40, 60)$. Table IV shows that the distortion of the level of significance of the score test (3) was negligible (also, see Appendix B). However, the empirical type 1 error rate of the Fleiss-like test (1) was

Table IV. Empirical type 1 error rates of homogeneity tests at $\alpha = 0.05$ for $\kappa_1 = \kappa_2 = \kappa$ derived under the assumption of $p_1 = p_2$ when $p_1 \neq p_2$ (based on 1000 simulations).

		$n_1 = 20, n_2 = 30$			$n_1 = 40, n_2 = 60$		
(p_1, p_2)	κ	X_F^2	X_G^2	X_s^2	X_F^2	X_G^2	X_s^2
(0.2, 0.3)	0.2	0.065	0.111	0.034	0.072	0.212	0.044
	0.4	0.107	0.117	0.043	0.073	0.193	0.050
	0.6	0.108	0.115	0.045	0.079	0.154	0.051
	0.8	0.095	0.102	0.039	0.080	0.139	0.038
(0.2, 0.5)	0.2	0.137	0.680	0.045	0.136	0.954	0.071
	0.4	0.165	0.633	0.046	0.170	0.925	0.080
	0.6	0.158	0.562	0.054	0.142	0.893	0.062
	0.8	0.129	0.497	0.024	0.142	0.842	0.048
(0.3, 0.5)	0.2	0.077	0.347	0.039	0.091	0.650	0.066
	0.4	0.081	0.301	0.039	0.080	0.552	0.042
	0.6	0.083	0.262	0.039	0.075	0.499	0.046
	0.8	0.075	0.231	0.029	0.074	0.452	0.037
(0.2, 0.8)	0.2	0.777	0.999	0.042	0.974	1.000	0.051
	0.4	0.659	0.999	0.045	0.914	1.000	0.052
	0.6	0.503	0.996	0.051	0.742	1.000	0.056
	0.8	0.384	0.993	0.031	0.496	1.000	0.042
(0.8, 0.2)	0.2	0.771	0.999	0.040	0.979	1.000	0.069
	0.4	0.674	0.999	0.041	0.904	1.000	0.073
	0.6	0.555	0.995	0.056	0.747	1.000	0.061
	0.8	0.403	0.990	0.032	0.520	1.000	0.044

X_F^2, X_G^2 and X_s^2 refer to the Fleiss-like (1), chi-square GOF (2) and score (3) tests.

substantially larger than the nominal level and that of the GOF test (2) was far greater. The p-values of the Fleiss-like and GOF tests were highly sensitive to violation of the assumption of equal prevalence while that of the score test was generally not affected. We may strongly recommend against use of the Fleiss-like test and the GOF test when the underlying assumption of equal prevalence is not appropriate.

6. EXAMPLE

Hannah *et al.* [19] presented data on alcohol drinking status by same-sex twins. The numbers of twin pairs by sex, zygosity and drinking status are shown in Table V.

For males, the prevalence of alcohol drinking was similar for monozygotic (MZ) and dizygotic (DZ) twins. The GOF and score tests for $\kappa_1 = \kappa_2$ yield p-values of 0.075 and 0.023, respectively. The difference between twin intraclass kappas, 0.462 and -0.033, was significant using the score test, but not by the GOF method. For females, the drinking prevalence was assumed equal for MZ and DZ twins. Under a model of equal prevalence, the GOF and score tests for comparing twin intraclass kappas, 0.474 and 0.360, gave p-values of 0.778 and 0.581. Unlike males, no significant difference between MZ and DZ twin intraclass kappas was found for females. Note that p-value of the GOF test was markedly different from that of the score test for both males and females.

Table V. Like-sex twin pairs by sex, zygosity and drinking status.

Alcohol drinking	Male			Female		
	MZ ($i=1$)	DZ ($i=2$)	Sum	MZ ($i=3$)	DZ ($i=4$)	Sum
Both	19	8	27	11	10	21
One	14	16	30	11	15	26
Neither	19	7	26	23	28	51
Sum	52	31	83	45	53	98
$\hat{p}_i$	0.500	0.516		0.367	0.330	
$\hat{\kappa}_i$	0.462	−0.033		0.474	0.360	

7. DISCUSSION

We have found that the homogeneity score test for kappa statistics derived under equal prevalence across studies is preferable to the Fleiss-like and GOF tests in terms of overall performance of accuracy of type 1 error rate and power for small or moderate sample sizes. The Fleiss-like test is anticonservative and the GOF test is less powerful than the score test. Prevalence in this paper means the probability of a positive rating and not the estimated positive rate. When the assumption of equal prevalence is invalid, the score test is very robust even when the difference among prevalences is extreme while the Fleiss-like and GOF tests are highly sensitive to a deviation from equal prevalence. If one use the Fleiss-like or GOF tests, the validity of the assumption of equal prevalence may be very important. Alternatively, one may apply the Donner's GOF procedure without assuming equal prevalence. However, a simulation study indicated that the test may be anticonservative under equal prevalence for small or moderate sample sizes.

If the homogeneity among kappa agreements in several studies is not rejected under the assumption of equal prevalence, we can pool all information into a single stratum and undertake inference on a summary kappa using the pooled data since prevalences are equal. If the homogeneity of kappa statistics is rejected, then we may proceed to further analyses to investigate the source of heterogeneity. When the equal prevalence assumption is not acceptable, the homogeneity tests for kappa statistics based on the assumption of unequal prevalences are more appropriate, e.g. References [11–13]. For this case, it has also been shown that the score test [13] is preferable over the other tests.

For some problems, it may be appropriate to assume equal prevalence across groups based on prior studies or theoretical justification, e.g. we have no reason to believe female MZ and DZ twins differ in drinking prevalence. For a very large sample size, a homogeneity test for kappa statistics using the equal prevalence provides more power than the corresponding test based on unequal prevalence when prevalences are the same.

Recently, the chance-corrected kappa agreement has been generalized in a number of ways, e.g. more than two raters and/or unbalanced data. Gonin *et al.* [20], Klar *et al.* [21], Williamson and Manatunga [22] and Thompson [23], have suggested use of regression models for transformed marginal probabilities or kappas, e.g. logit transformation of marginal probabilities or Fisher's z-transformed weighted kappa, and proposed generalized

estimating equation for the regression parameters. When homogeneity of kappa statistics across groups is rejected, researchers want to know sources of heterogeneity, and the general modelling of weighted kappa may be very useful to determine whether the differences in agreement could be related to other variables. Since a probabilistic model that generates the data varies from problem to problem, it is very important to specify correctly a valid model for a given problem. It is prudent to examine the appropriateness of a model before conducting a regression analysis. Although estimating equations may not be often efficient, they have performed well in many cases. The homogeneity tests and inference on kappa agreement in this paper are based on a straightforward multinomial probability model for observed data, and general modelling of weighted kappa statistics is not within the scope of this investigation.

We should exercise a careful judgment regarding whether or not comparing reliabilities appropriate for a given problem, e.g. comparing the MZ *versus* DZ twin intraclass kappas for the same sex, MZ for males *versus* MZ for females or DZ for males *versus* DZ for females are appropriate, but comparing the male MZ kappa *versus* the female DZ kappa, or the male DZ kappa *versus* the female MZ kappa would not make sense and the homogeneity test should not used for this case.

The normality assumption of a kappa statistic and estimated large-sample variance of the kappa statistic have been often used in the derivation of statistical methods related to inference of kappa agreement, e.g. Reference [11]. The homogeneity test for kappa statistics based on this approach is intuitive and simple. However, convergence of the distribution of a kappa statistic to the normal is very slow with respect to sample size. As indicated by a simulation study, the Fleiss-like test is seriously biased and its p-values are substantially different from nominal level for small or moderate sample sizes. We recommend strongly against application of the Fleiss-like test and also sample size determination using this test unless the sample size is very large.

APPENDIX A: THE jTH SCORE EVALUATED AT $\kappa_j = \hat{\kappa}$ AND $p = \hat{p}$

The jth score under $p_j = p$ is expressed as

$$\begin{aligned} S_j(\kappa_j, p) &\equiv \partial \ln L_j(\kappa_j, p)/\partial \kappa_j \\ &= x_{2j}q/(p + q\kappa_j) - x_{1j}/(1 - \kappa_j) + x_{0j}p/(q + p\kappa_j) \\ &= \{x_{2j}q(1 - \kappa_j)/(p + q\kappa_j) - x_{1j} + x_{0j}p(1 - \kappa_j)/(q + p\kappa_j)\}/(1 - \kappa_j) \\ &= \{x_{2j}(1 - p - q\kappa_j)/(p + q\kappa_j) - x_{1j} + x_{0j}(1 - q - p\kappa_j)/(q + p\kappa_j)\}/(1 - \kappa_j) \\ &= \{x_{2j}/(p + q\kappa_j) + x_{0j}/(q + p\kappa_j) - n_J\}/(1 - \kappa_j) \quad \text{since } n_j = x_{2j} + x_{1j} + x_{0j} \end{aligned}$$

Thus, the jth score evaluated at $\kappa_j = \hat{\kappa}$ and $p = \hat{p}$ is written as

$$\{S_j(\kappa_j, p)\}_{\kappa_j = \hat{\kappa},\, p = \hat{p}} \equiv S_j(\hat{\kappa}, \hat{p}) = \{x_{2j}/(\hat{p} + \hat{q}\hat{\kappa}) + x_{0j}/(\hat{q} + \hat{p}\hat{\kappa}) - n_j\}/(1 - \hat{\kappa}) \tag{A1}$$

APPENDIX B: TRUE LEVEL OF SCORE TEST AT α UNDER UNEQUAL PREVALENCE

Consider the jth likelihood score and its variance evaluated at $\kappa_j = \hat{\kappa}$ and $p = \hat{p}$ for $j = 1, 2, \ldots, J$. As n_j increases, the score, (A1), and its variance under $p_j \neq p$ approach their respective asymptotic values, i.e.

$$S_j(\hat{\kappa}, \hat{p}) \to n_j \left\{ \frac{p_j(p_j + q_j\kappa)}{\bar{p} + \bar{q}\kappa} + \frac{q_j(q_j + p_j\kappa)}{\bar{q} + \bar{p}\kappa} - 1 \right\} \Big/ (1 - \kappa) \quad \text{and} \quad \omega(\hat{\kappa}, \hat{p}) \to n_j / v$$

where $v = (1 - \kappa)^2[(1 - 2\kappa) + \kappa(2 - \kappa)/\{2pq(1 - \kappa)\}]$, $\bar{p}$ is the asymptotic value of $\hat{p}$ and $\bar{q} = 1 - \bar{p}$. When $p_j \neq p$, the score statistic, (3), is distributed asymptotically as a non-central chi-square with $J - 1$ degrees of freedom and non-centrality parameter

$$\lambda'_s = c \left[\sum_{j=1}^{J} t_j \left\{ \frac{p_j(p_j + q_j\kappa)}{\bar{p} + \bar{q}\kappa} + \frac{q_j(q_j + p_j\kappa)}{\bar{q} + \bar{p}\kappa} - 1 \right\}^2 \right] n. \tag{B1}$$

where $c = 1 - 2\kappa + \kappa(2 - \kappa)/\{2pq(1 - \kappa)\}$, $t_j = n_j/n.$ and $n. = \sum_{j=1}^{J} n_j$.

When prevalences are unequal, the true value of the homogeneity score test for kappa statistics at α using (3) is

$$\Pr(X_s^2 \geqslant \chi^2_{J-1,1-\alpha} | p_j \neq p) = \Pr\{\chi^2_{J-1}(\lambda'_s) \geqslant \chi^2_{J-1,1-\alpha}\} \tag{B2}$$

where $\chi^2_{J-1,1-\alpha}$ is the $100 \times (1 - \alpha)$ percentile point of the chi-square distribution with $J - 1$ degrees of freedom and $\chi^2_{J-1}(\lambda'_s)$ is the non-central-chi-square distribution with $J - 1$ degrees of freedom and non-central parameter λ'_s. Form (B2) shows that the distortion of the p-value of the test is a monotone increasing function of λ'_s. If λ'_s is very small then the difference between true and nominal α of the score test is negligible.

Suppose $p_j(p_j + q_j\kappa)/(\bar{p} + \bar{q}\kappa) > p_j$ which is equivalent to $p_j + q_j\kappa > \bar{p} + \bar{q}\kappa$ since $p_j > 0$. By replacing $p_j = 1 - q_j, q_j = 1 - p_j, \bar{p} = 1 - \bar{q}$ and $\bar{q} = 1 - \bar{p}$ in the above inequality, we have $q_j + p_j\kappa < \bar{q} + \bar{p}\kappa$ or $q_j(q_j + p_j\kappa)/(\bar{q} + \bar{p}\kappa) < q_j$ for $j = 1, 2, \ldots, J$. If the first term inside the curly brackets is larger than p_j, then the following second term is smaller than q_j. Similarly, if the first term is smaller than p_j, then the second term is larger than q_j. The two terms are counter-balancing toward unity since $p_j + q_j = 1$. Numerical evaluation shows that the non-centrality parameter, (B1), is close to zero unless the total sample size is extremely large. For typical sample sizes used in reliability studies, the distortion of the level of significance of the score test when prevalences are unequal, (B2), is negligible.

ACKNOWLEDGEMENTS

The author is very grateful to two anonymous referees for their helpful suggestions and constructive comments on an earlier version of the article.

REFERENCES

1. Cohen J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 1960; **20**:37–46.
2. Scott WA. Reliability of content analysis: the case of nominal scale coding. *Public Opinion Quarterly* 1955; **19**:321–325.

3. Bloch DA, Kraemer HC. 2 × 2 kappa coefficients: measure of agreement or association. *Biometrics* 1989; **45**: 269–287.
4. Kraemer HC, Periyakoil VS, Noda A. Tutorial in biostatistics: kappa coefficients in medical research. *Statistics in Medicine* 2002; **21**:2109–2129.
5. Dunn G. *Design and Analysis of Reliability Studies*. Oxford University Press: New York, 1989.
6. Wright S. The genetical structure of populations. *Annals of Eugenics* 1951; **15**:322–354.
7. Donner A, Eliasziw M. A goodness-of-fit approach to inference procedures for the kappa statistic: confidence interval construction, significance-testing and sample-size estimation. *Statistics in Medicine* 1992; **11**:1511–1519.
8. Hale CA, Fleiss JL. Interval estimation under two study designs for kappa with binary classifications. *Biometrics* 1993; **49**:523–524.
9. Nam J. Interval estimation of the kappa coefficient with binary classification and an equal marginal probability model. *Biometrics* 2000; **56**:583–585.
10. Nam J. Testing the intraclass version of kappa coefficient of agreement with binary scale and sample size determination. *Biometrical Journal* 2002; **44**:558–570.
11. Fleiss JL. *Statistical Methods for Rates and Proportions.* Wiley: New York, 1981.
12. Donner A, Eliasziw M, Klar N. Testing the homogeneity of kappa statistic. *Biometrics* 1996; **52**:176–183.
13. Nam J. Homogeneity score test for the intraclass version of the kappa statistics and sample size determination in multiple or stratified studies. *Biometrics* 2003; **59**:1027–1035.
14. Mak TK. Analyzing intraclass correlation for dichotomous variables. *Applied Statistics* 1988; **37**:344–352.
15. Johnson NL, Kotz S. *Continuous Univariate Distributions-2*, Chapter 28. Houghton Mifflin: Boston, MA, 1970.
16. Haynam GE, Govindarajulu Z, Leone GC. Tables of the cumulative non-central chi-square distribution. *Case Statistical Laboratory Publication No. 104*. Part of the tables have been published in *Selected Tables in Mathematical Statistics*, Harter HL, Owen DB (eds), vol. 1, 1962.
17. Rao CR. *Linear Statistical Inference and its Application.* Wiley: New York, 1973.
18. Bera AK, Bilias Y. Rao's score, Neyman's C(alpha) and Silvey's LM tests: an essay on historical developments and some new results. *Journal of Statistical Planning and Inference* 2001; **97**:9–44.
19. Hannah MC, Hopper JL, Mathews JD. Twin concordance for a binary trait. I. Statistical models illustrated with data on drinking status. *Acta Geneticae Medicae et Gemellologiae* 1983; **32**:127–137.
20. Gonin R, Lipsitz, Fitzmaurice GM, Molenberghs G. Regression modelling of weighted κ by using generalized estimating equations. *Applied Statistics* 2000; **49**:1–18.
21. Klar N, Lipsitz SR, Ibrahim J. An estimating equation approach for modelling kappa. *Biometrical Journal* 2000; **42**:45–58.
22. Williamson JR, Manatunga AK. Assessing interrater agreement from dependent data. *Biometrics* 1997; **53**: 707–714.
23. Thompson JR. Estimating equations for kappa statistics. *Statistics in Medicine* 2001; **20**:2895–2906.